SAND2014-16120R

SANDIA NATIONAL LABORATORIES

Analysis of Spectral Variation in Background Gamma Radiation at U.S. Ports of Entry

Kaylee L. Homolka

Summer 2014

I was part of the Technical Reachback (TRB) program at the Sandia National Laboratories sponsored by the Domestic Nuclear Detection Office of the Department of Homeland Security. This project exists to improve the ability to detect nuclear material crossing the borders. The purpose of my portion of this project was to analyze spectral variations of backgrounds measured by radiation portal monitor detectors at ports of entry using available data to determine whether or not the measured background variation includes changes in the spectral shape in addition to changes in the gross counts. I analyzed background gamma-ray measurements that can be modeled as a multidimensional vector, called a spectrum. The spectrum is composed of measured counts of the detected gamma rays in 8 energy bins, or energy windows. The 8 energy windows in the spectrum can be represented by the vector $\boldsymbol{B} = [B_1, B_2, \dots, B_8]$ where B_i is the i^{th} window. The statistical analysis I performed on the spectrum is used to validate the relationship between the 8 energy windows measured by the detector. Large deviations from the predicted linear behavior could be due to real spectral changes or to a malfunctioning radiation detector. The analysis I performed also checks for a constant spectral shape, meaning the expected linear behavior holds within the physically present statistical uncertainty caused by Poisson noise.

The results of this project can be used to aid in the setting of gross counts and spectral alarm thresholds, for optimizing detection algorithms, for verifying that the radiation portal is not malfunctioning, for detecting inconsistencies in the background readings, and for discovering physical spectral changes in the background. Some feasible causes for background variation include changes in temperature resulting from the ambient weather conditions, physical changes at the port of entry (like construction on the roads), the presence of radioactive sources near the

radiation portal monitor (either in equipment or facilities nearby), or changes in weather that might alter the radon concentration. This study aimed at statistically analyzing this variation.

To begin, I had to acquire a proper data set from the provided database. This included cross checking tables and verifying software versions for the radiation portal monitors for a selected port of entry in the database. For this task I had to research and learn a few basic SQL commands. I chose one location (one port of entry) to perform the analysis on. The resulting validated data set contains background gamma measurements for multiple radiation sensor panels (RSP) in the designated port of entry. The analysis that followed was implemented on one RSP at a time since the background readings can vary between the panels due to differences in their physical location and orientation.

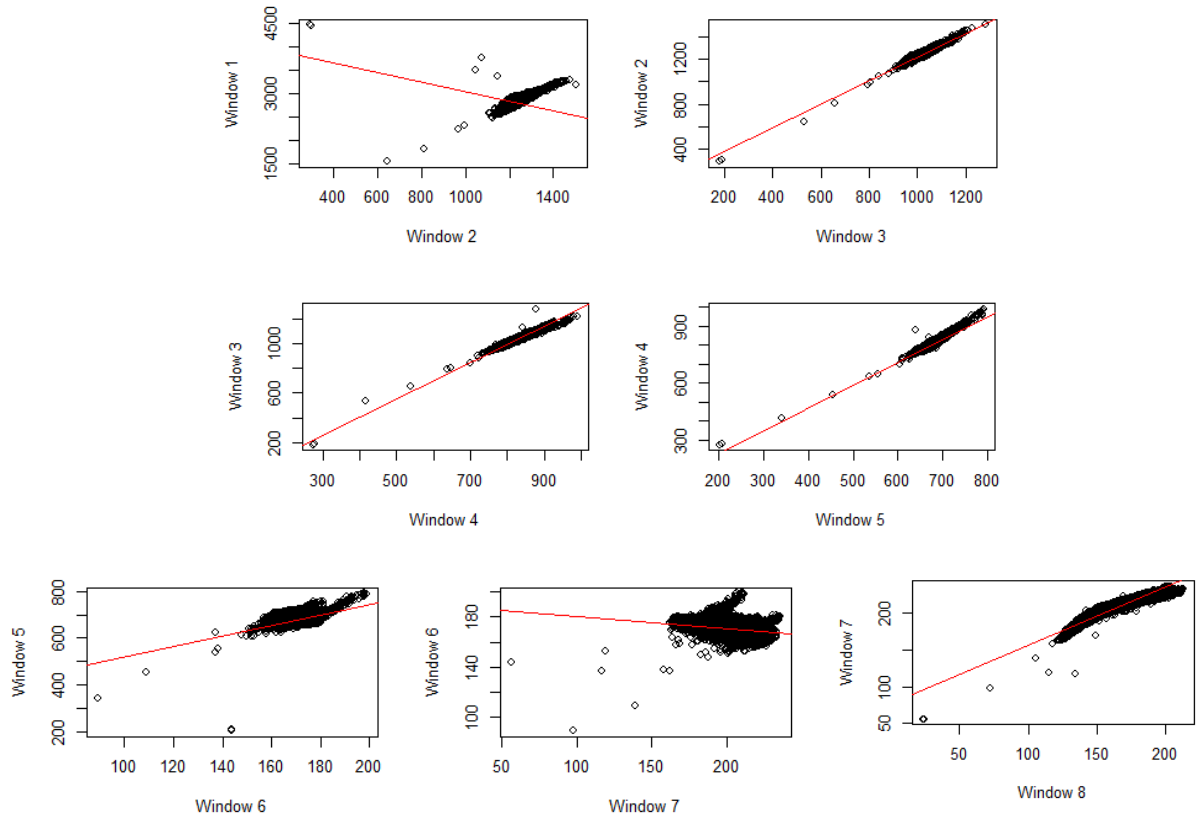
Once a data set was verified by comparing different data entries in the database, the Poisson noise level was estimated. The background counts for the radiation sensor panels are given in units of counts per second, but a check on the averaging period was needed. This was estimated using approximately 100 data points that were selected so that it could be expected that background variation across the points was due to statistical noise and not a physical effect. I tested several sets of points throughout the 8 energy windows. For example, I selected the data from points 200 to 300 in window 1 from the first RSP. The noise level was estimated using root mean square deviation about the mean which is given by $N = \{\frac{1}{n}\sum_{i=1}^{n}[B_i - < B >]^2\}^{1/2}$ where N is the estimated noise level, n is the number of samples in the dataset, B_i is the background for the i^{th} sample, and $< B >$ is the mean background value for the n samples. I performed these calculations in R Studio, and for window 1 from points 200 to 300 I found that $N = 13.6$. N

gives an estimate of the background radiation due to statistical noise. Higher variance is therefore expected to be due to physical changes of the detector or background.

I then used the value of N to calculate the averaging period, τ , in seconds. Since the total background counts over the period is $< B > \tau$ and the statistical Poisson noise is $\sqrt{< B > \tau}$, the noise for the given background in the data is approximately $\sqrt{(< B > \tau)}/(\tau)$. Then τ can be estimated by $N = \sqrt{(< B > \tau)}/(\tau)$ using N from the original equation. Performing this calculation on the selected points from window 1 I found $\tau = 15.2489$ seconds. I expected an averaging period between 1 and 60 seconds, and for the dataset I analyzed it checked out. I tested several portions of data from varied windows of all of the RSPs for the node I chose to analyze. All of them behaved as expected.

Following the verification of the Poisson noise level estimation, I progressed to analyzing the individual background windows. Changes in the background that only alter the length of the spectral vector do not impact the spectral shape. The spectral variation study was used to verify this.

To begin I tested if background variation at the port of entry is only in overall counts, meaning only the length of the spectral vector changes. With this assumption and ignoring statistical noise, the background at time t can be written as $\boldsymbol{B}(t) = \alpha(t)\boldsymbol{B}_m$ where $\boldsymbol{B}_m$ is the mean background spectral vector and $\alpha(t)$ is a scalar that depends on time. Then it follows that plots of one window versus another window should be linear (with zero intercept) and thus their correlation should be equal to unity, but this is not the exact case due to statistical noise. I tested this hypothesis visually and by calculation. To test visually, I plotted the windows against one another and looked for linearity.



The plot for window 1 versus window 2 has a negative sloping line due to the shape of the data plotted. It appears linear, but when it is plotted on a smaller scale it can be seen that the line fits the present data. The other plots appear to be generally linear with the exception of window 6 versus window 7, but there are outliers that could be skewing the fit or there could be spectral changes. The low correlation between some of the windows (1 vs 2, 5 vs 6, and 6 vs 7) was of further interest.

I then used linear regression to fit the data to the equation $B_i = aB_j + b$, where B_i is window i counts, a and b are constant coefficients, and B_j is window j counts. For windows i and j, we get $B_i(t) = \alpha(t)\, B_{m-i}$ and $B_j(t) = \alpha(t)\, B_{m-j}$, which gives $B_i(t) = C\, B_j(t)$ where C is a constant $C = B_{m-i}/B_{m-j}$ and t is time. This is true only if the data is truly linear, so then I used linear regression to verify linearity. I used the results from the equation $B_i(t) = aB_j(t) + b$

where a and b are calculated using least squared minimization. This is done by minimizing the chi-squared residual $\chi^2 = \sum_{k=1}^{n} \frac{\left[B_i(t_k) - \left(a\, B_j(t_k) + b\right)\right]^2}{\left[\sigma(t_k)\right]^2}$ where $\left[\sigma(t_k)\right]^2$ is the variance of the measurement at time t_k. It follows that $B_i(t_k)$ and $B_j(t_k)$ can be used to estimate their respective variances and $\left[\sigma(t_k)\right]^2 = B_i(t_k) + a^2 B_j(t_k)$ is the overall variance. I used R Studio to calculate the normalized chi-squared residual (multiply the chi-squared residual by $1/n$). For example, using windows 1 and 2 the chi-squared residual was calculated as chi-squared= $\sum_{k=1}^{55571}[Win1 - (-1.0240 * Win2 + 4056.5200)]^2/[Win1 + (-1.0240^2 * Win2)] = 177733$ (note that $n = 55571$ since that is how many points were in the data set for window 1 at this RSP). Next this was divided by n, $\frac{177733}{55571} = 3.1983$. The results are shown in the table below.

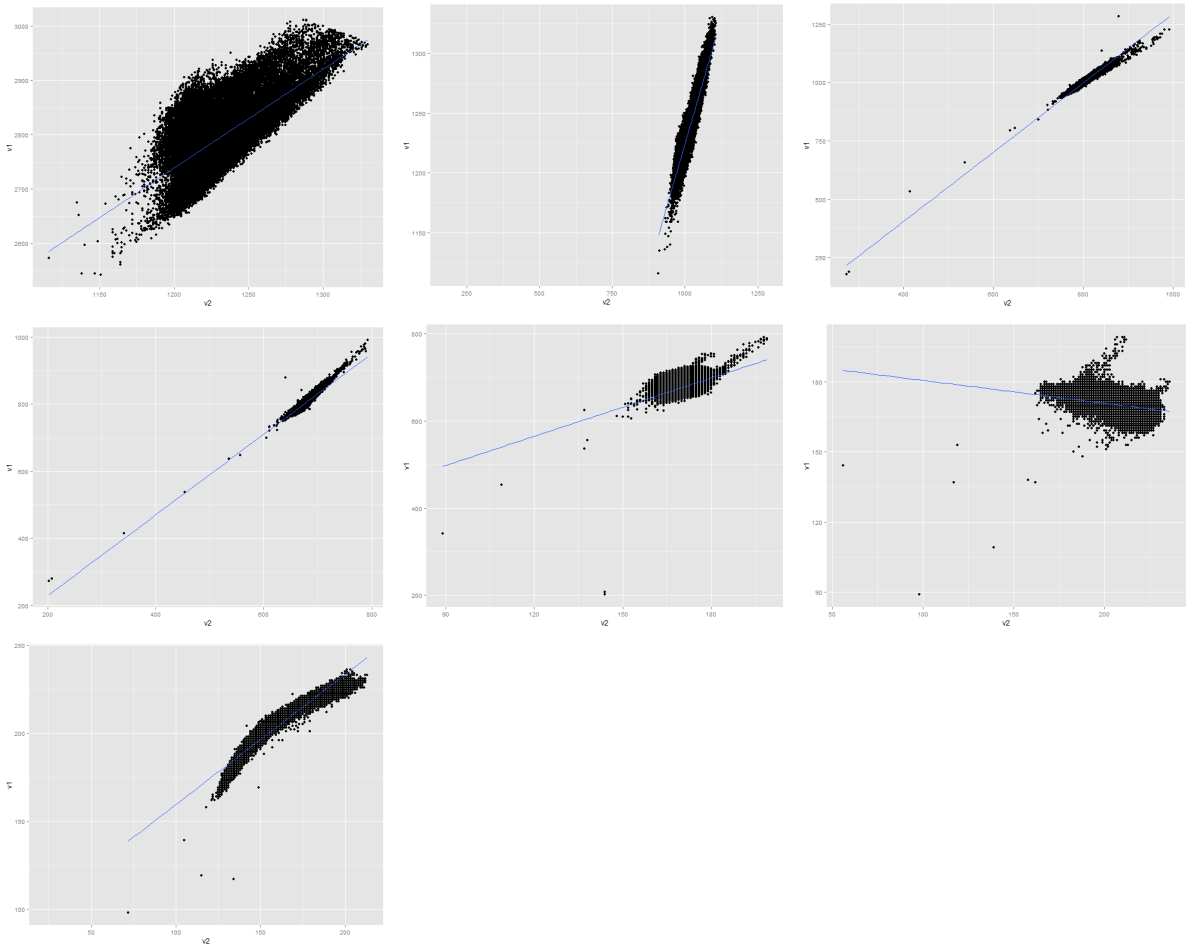
$B_i(t_k) = a$ $B_j(t_k) + b$	X^2	Adj. R^2	Std. Dev a	Std. Dev b	Cor.
Win1=-1.0240* Win2+4056.5200	3.1983	0.2597	0.0073	9.0180	-0.5096
Win2=1.0550* Win3+167.9560	1.0653	0.9655	0.0008	0.8505	0.9826
Win3=1.4840* Win4-188.3080	1.0006	0.9844	0.0008	0.6384	0.9922
Win4=1.2040* Win5-11.6980	1.0354	0.9554	0.0011	0.7488	0.9774
Win5=2.2270* Win6+297.6100	1.6647	0.1206	0.0255	4.3569	0.3472
Win6=-0.0979* Win7+190.3684	1.1044	0.1199	0.0011	0.2269	-0.3464
Win7=0.7891* Win8+77.7260	1.1035	0.9306	0.0009	0.1441	0.9647

This table contains the results for the first RSP. The first column is the linear regression. The second column contains the normalized chi-squared residual. If this is too large, then the hypothesis is not valid and the data set contains more than one spectral shape. To further analyze data with a large chi-squared value, these calculations can be performed on subdivided data. If

the chi-squared residual is reasonable and the intercept, b, is close to zero, then the data set represents one spectral shape within the statistical uncertainty due to noise. In this case, b is not close to zero for any of the windows. The adjusted R^2 value estimates how much the variance in the data is explained by the linear model. The standard deviations of a and b are estimates from R Studio. The confidence intervals of a and b were also calculated in R Studio, and both a and b are in a 95% confidence interval. For example, for windows 1 and 2 for a to be in a 95% confidence interval it must be between -1.0387 and -1.0099 and b must be between 4038.8446 and 4074.1948. Since a and b are in these values, they fit in the 95% confidence interval (this is done using the 'confint' function in R Studio). The correlation was also calculated in R Studio.

I then attempted to isolate anomalies and outliers to further analyze the data. I used a script in R Studio to plot the measurements before outlier removal with the linear regression line, then remove the outliers and plot the new results. To do this, I took two windows and assumed they would fit the model $W_2(j) = aW_1(j) + b$ where j is the background measurement. To get coefficients a and b I used least squares analysis without variance weighting (variance=1.0 for all). Then I found the root mean squared deviation (rmsd) between the data and the linear fit, where $\text{RMS Deviation} \equiv D_{RMS} = \left\{ \frac{1}{n} \sum_{k=1}^{n} \left[W_2(k) - \left(aW_1(k) + b \right) \right]^2 \right\}^{1/2}$. Then for the j[th] background measurement, the deviation is $D_j = \left\{ \left[W_2(j) - \left(aW_1(j) + b \right) \right]^2 \right\}^{1/2}$ which is the absolute value of the deviation. Then any measurement where D_j is larger than a chosen multiple (I used 3) of D_{RMS} can be considered an outlier. After these outliers are removed linear regression analysis can be repeated. The results were as follows.

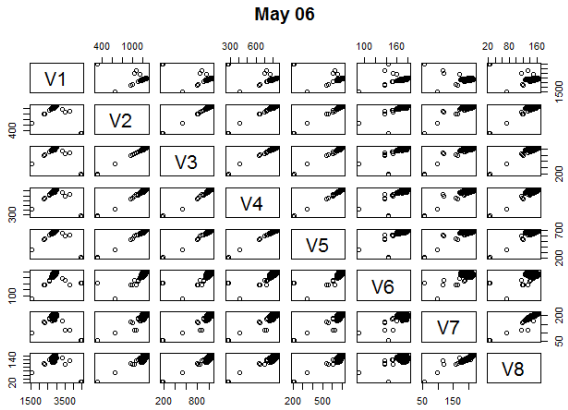
$B_i = aB_j + b$
Win1=1.8220*Win2+552.3820
Win2=0.8380*Win3+386.1900
Win3=1.4840*Win4-188.3080
Win4=1.2040*Win5-11.6980
Win5=2.2270*Win6+297.6100
Win6=-0.1885*Win7+208.7605
Win7=0.8100*Win8+75.2360



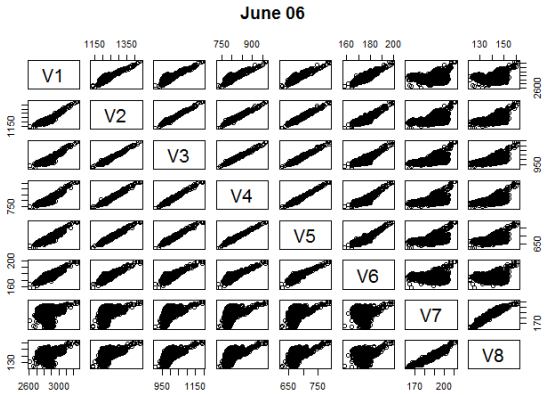
Using this outlier removal method caused the regression line to fit the data closer.

To further analyze the data, I divided it into several sets that could be analyzed independently. I chose to divide the data set by month. I used Microsoft Excel to divide the data, and loaded the separated data into R Studio. I tested the data for each month visually and with linear correlation like I did for the complete data set. I observed the plots above the diagonal in

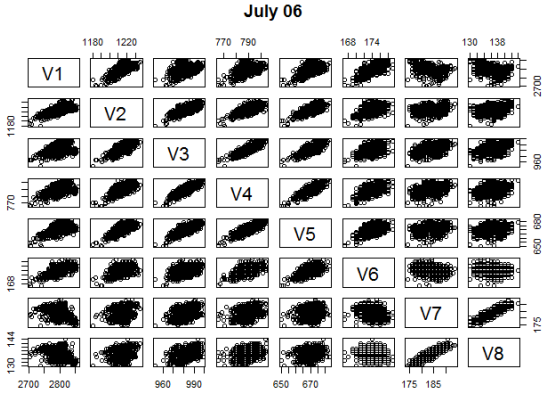
the following diagrams, and I used R studio to calculate linear regression coefficients and the correlation.



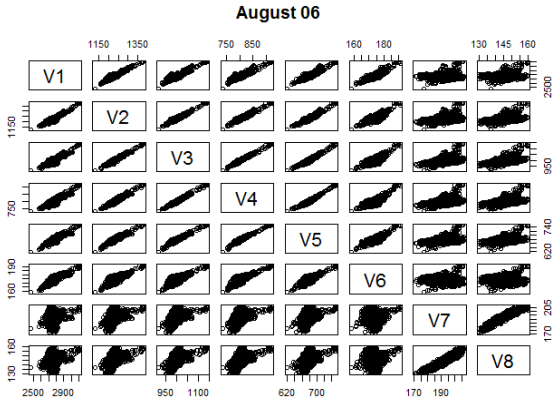
Linear Regression	Cor
Win1=-1.7120*Win2+4921.6800	-0.9608
Win2=1.1380*Win3+88.8750	0.9983
Win3=1.5490*Win4-239.0100	0.9985
Win4=1.1120*Win5+46.7790	0.9981
Win5=10.8300*Win6-1190.1700	0.8173
Win6=0.1507*Win7+141.7682	0.6312
Win7=1.1360*Win8+30.6720	0.9945



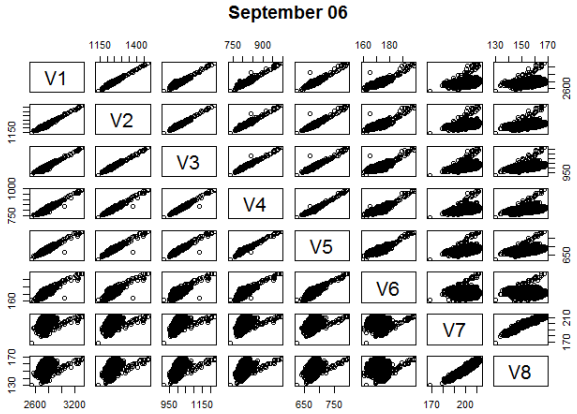
Win1=1.9640*Win2+430.3310	0.7581
Win2=0.9709*Win3+260.0606	0.9510
Win3=1.1560*Win4+69.8050	0.9646
Win4=1.3890*Win5-141.7130	0.9436
Win5=2.2690*Win6+277.6050	0.6923
Win6=-0.1072*Win7+194.1659	-0.2647
Win7=1.2890*Win8+7.7440	0.9643



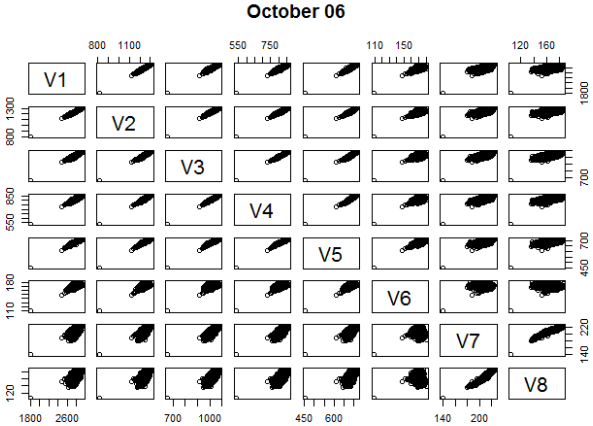
Win1=2.4260*Win2-147.4770	0.6816
Win2=0.7491*Win3+478.4381	0.7242
Win3=0.9967*Win4+195.467	0.8523
Win4=1.0210*Win5+104.4210	0.8297
Win5=1.6230*Win6+387.8120	0.6181
Win6=-0.1296*Win7+198.0644	-0.2711
Win7=1.3150*Win8+4.2580	0.8958



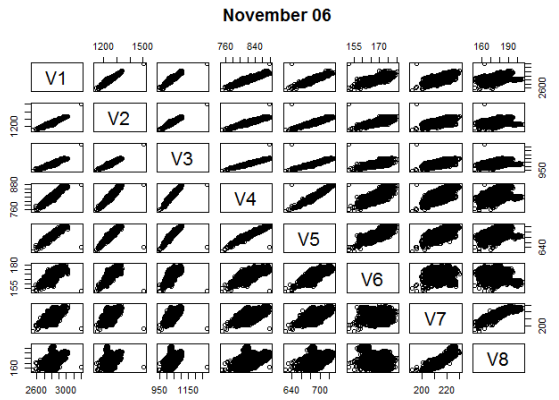
Win1=1.9520*Win2+368.1800	0.8245
Win2=0.9561*Win3+271.5613	0.9314
Win3=1.2130*Win4+26.3290	0.9615
Win4=1.4110*Win5-155.8350	0.9377
Win5=2.0100*Win6+325.8700	0.6230
Win6=-0.1274*Win7+197.2704	0.3259
Win7=1.1600*Win8+24.5000	0.9557



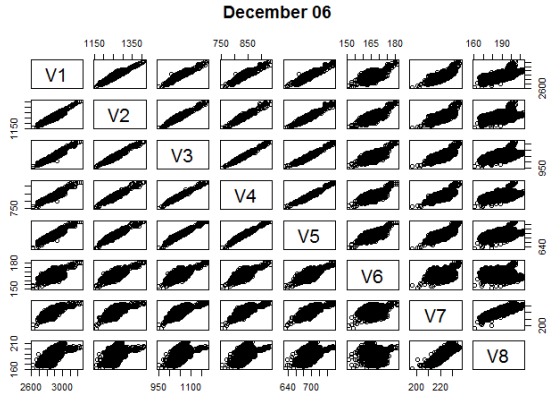
Win1=2.2020*Win2+39.2000	0.9373
Win2=1.0050*Win3+219.8450	0.9475
Win3=1.2120*Win4+29.8940	0.9694
Win4=1.4440*Win5-176.2920	0.9504
Win5=2.2510*Win6+290.0220	0.7099
Win6=-0.1481*Win7+200.9173	-0.3027
Win7=0.9550*Win8+53.9240	0.9548



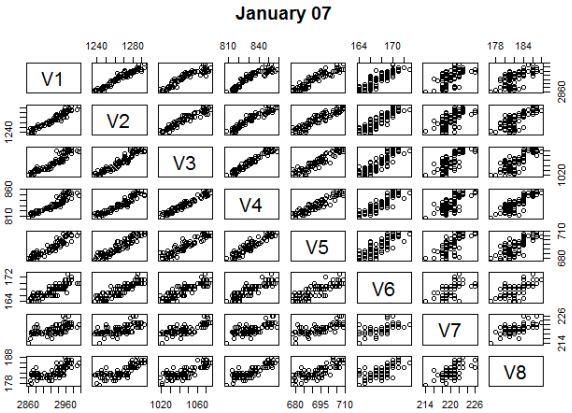
Win1=2.2750*Win2-48.4560	0.9614
Win2=0.9632*Win3+257.7319	0.9609
Win3=1.2920*Win4-32.0340	0.9772
Win4=1.4130*Win5-152.5770	0.9570
Win5=2.0000*Win6+341.4000	0.6005
Win6=-0.0998*Win7+190.2250	-0.2269
Win7=0.7696*Win8+82.1163	0.9676



Win1=2.1010*Win2+186.2390	0.9489
Win2=1.0270*Win3+191.2980	0.9545
Win3=1.2660*Win4-9.5280	0.9802
Win4=1.4280*Win5-160.6360	0.9675
Win5=2.3850*Win6+284.9220	0.7244
Win6=-0.0690*Win7+183.8175	-0.0959
Win7=0.5365*Win8+121.5191	0.9261



Win1=2.2470*Win2+52.5920	0.9185
Win2=1.0440*Win3+166.9250	0.9586
Win3=1.2820*Win4-20.6510	0.9812
Win4=1.4100*Win5-146.4700	0.9713
Win5=2.5340*Win6+269.1210	0.6458
Win6=0.0649*Win7+151.6742	0.0956
Win7=0.4455*Win8+137.6698	0.8922



Win1=2.3140*Win2-1.8080	0.9599
Win2=0.8736*Win3+351.5599	0.9428
Win3=1.1830*Win4+62.2650	0.9700
Win4=1.3870*Win5-129.7550	0.9363
Win5=3.3490*Win6+131.4950	0.8134
Win6=0.5849*Win7+39.0311	0.5737
Win7=0.5816*Win8+113.8630	0.6636

It is important to note that July and January were missing some data in the database. The results of these tests on the divided data generally matched the results of the overall calculations: all windows ended up being relatively highly correlated except for window 6 versus window 7.

There were a few exceptions where other windows had lower correlations, but upon further analysis it appeared that these were due to extreme outliers. An example of this is for May 2006 windows 1 and 2 were negatively correlated, but with the removal of outliers the correlation increased. This analysis could have been done with any time division (like by day) or by clustering, but I found by month to be the most straightforward way.

I ran these statistics on the first RSP divided by month. The results were similar to the table for the overall RSP. I also calculated these statistics for all of the other RSPs on the port I had selected. The results were very similar to the first RSP. Based on these statistics I could conclude that the data for all of the RSPs of the port I chose does not represent one spectral shape due to the large values for the intercept b. These results can help in setting gross counts and alarm thresholds, in optimizing algorithms, in verifying the functionality of the radiation portals, in the detection of inconsistencies in background readings, and in discovering physical spectral changes.

This project gave me a good experience of what math is like outside of the classroom, and it gave me a new view on the real world application of mathematics. Now that I have some experience of math beyond school, I can better prepare myself for graduation and my future career by careful selection of my upcoming courses. This internship was the experience I needed to narrow my career planning and focus my academics.

I have always been asked what kind of career I plan on having with a major in mathematics, and I have never had an idea as to what people usually do with it once they graduate. I had the opportunity to meet with several people during my time on this project, many of them with degrees in math, and they had plenty of insight to offer on this matter. I also had the chance to attend several lectures on a wide variety of topics, including math and computer science. Not only were the talks informative and educational in their content, but they were also beneficial to watch to learn how to communicate scientific ideas clearly to an audience. The knowledge of the other employees combined with the talks I attended greatly impacted my academic and career preparation.

This project also provided me with many new challenges. I had to work to expand my knowledge of statistics, and I enjoyed the task. This research has given me an opportunity to focus on an area of math that I am not extremely familiar with and work towards a greater understanding. A large portion of the statistics I used in this analysis was beyond the scope of the statistics course I had taken in college, and it was interesting to research those subjects on my own. I appreciated the freedom that came with this project. My mentor provided guidelines and an overview of the project and gave me the ability to explore and test the data however I could think to do so. This greatly promoted my critical thinking and problem solving skills. My mentor

was great and always ready to help whenever I needed it. Overall this internship was a significant benefit to me and my future. Thank you for the opportunity.